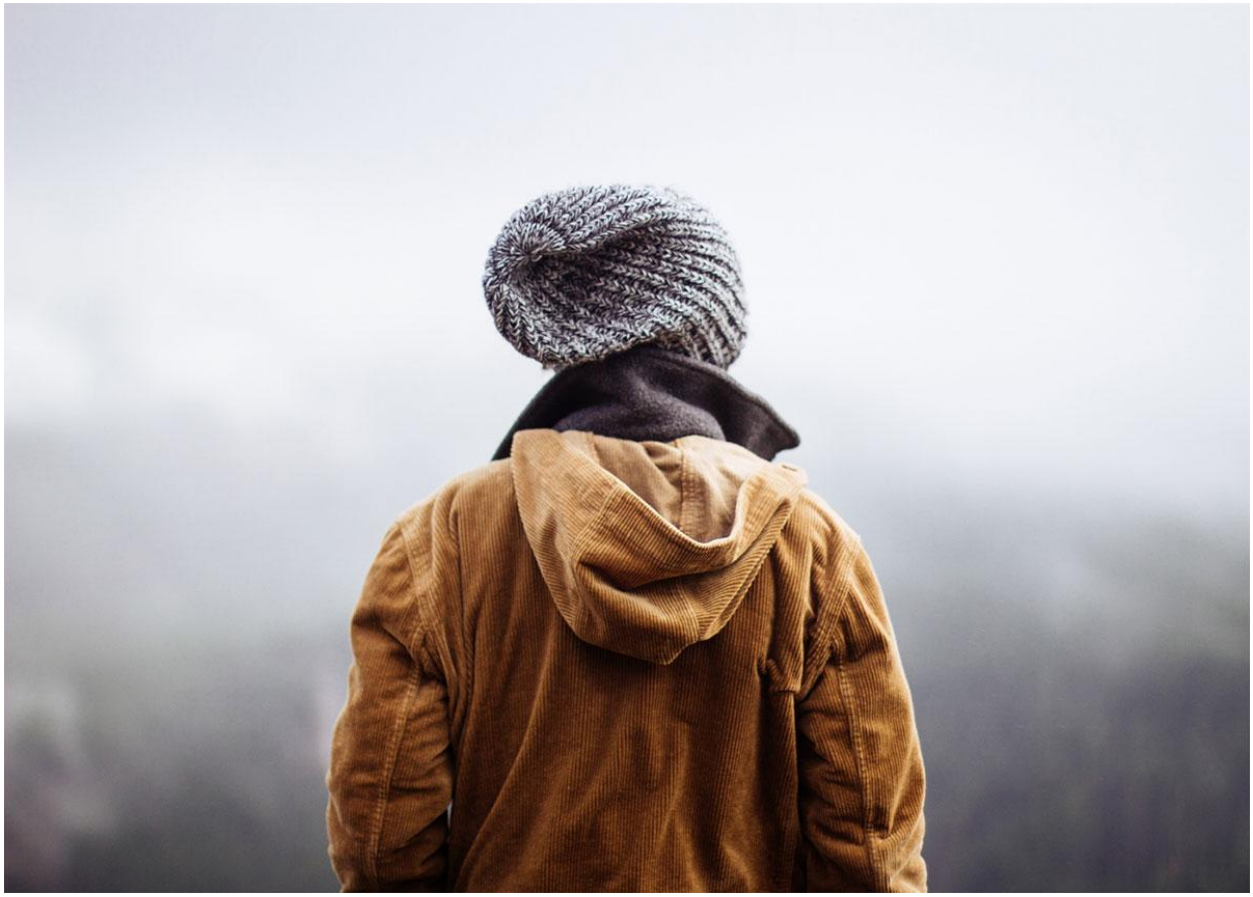
Tan Khuu

Lead Scoring Case Study

Model Predict a Customer is a HotLead or not



Introduction

```
# Lead Scoring in X Education
```

The company wishes to identify the most potential leads, also known as 'Hot Leads'.

If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

There are quite a few goals for this case study:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Results Expected

A well-commented Jupyter notebook with at least the logistic regression model, the conversion predictions and evaluation metrics.

The word document filled with solutions to all the problems.

The overall approach of the analysis in a presentation.

Mention the problem statement and the analysis approach briefly

Explain the results in business terms

Include visualisations and summarise the most important results in the presentation

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

You need to submit the following four components:

- Python commented file: Should include detailed comments and should not contain unnecessary pieces of code.

- Word File: Answer all the questions asked by the company in the word document provided.

- Presentation: Make a presentation to present your analysis to the chief data scientist of your company (and thus you should include both technical and business aspects). The presentation should be concise, clear, and to the point. Submit the presentation after converting it into PDF format.

- PDF File: Write the summary report in a word file and submit it as a PDF.

 # Model Predict a Customer is a HotLead or not

ROC Curve is close to the left hand and near on the top => It's good

Making Predictions

#scaling test set

num_cols=X_test.select_dtypes(include=['float64', 'int64']).columns X_test[num_cols] = scaler.fit_transform(X_test[num_cols]) X_test.head() X_test = X_test[rfe_columns] X_test.head() X_test_sm = sm.add_constant(X_test) y_test_pred = res.predict(X_test_sm) y_test_pred[:10]

Converting y_pred to a dataframe which is an array

y_pred_1 = pd.DataFrame(y_test_pred)

Let's see the head

y_pred_1.head()

Converting y_test to dataframe

y_test_df = pd.DataFrame(y_test)

Putting CustID to index

y_test_df['Prospect ID'] = y_test_df.index

Removing index for both dataframes to append them side by side

y_pred_1.reset_index(drop=True, inplace=True) y_test_df.reset_index(drop=True, inplace=True)

Appending y_test_df and y_pred_1

y_pred_final = pd.concat([y_test_df, y_pred_1],axis=1) y_pred_final.head()